BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 19 2014, pages 2702–2708
doi:10.1093/bioinformatics/btu369

Genome analysis

Advance Access publication June 4, 2014

Improving translation initiation site and stop codon recognition by using more than two classes

Javier Pérez-Rodríguez, Alexis G. Arroyo-Peña and Nicolás García-Pedrajas*

Department of Computing and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, Edificio Einstein, Planta 3, 14071 Córdoba, Spain

Associate Editor; John Hancock

ABSTRACT

Motivation: The recognition of translation initiation sites and stop codons is a fundamental part of any gene recognition program. Currently, the most successful methods use powerful classifiers, such as support vector machines with various string kernels. These methods all use two classes, one of positive instances and another one of negative instances that are constructed using sequences from the whole genome. However, the features of the negative sequences differ depending on the position of the negative samples in the gene. There are differences depending on whether they are from exons, introns, intergenic regions or any other functional part of the genome. Thus, the positive class is fairly homogeneous, as all its sequences come from the same part of the gene, but the negative class is composed of different instances. The classifier suffers from this problem. In this article, we propose the training of different classifiers with different negative, more homogeneous, classes and the combination of these classifiers for improved accuracy.

Results: The proposed method achieves better accuracy than the best state-of-the-art method, both in terms of the geometric mean of the specificity and sensitivity and the area under the receiver operating characteristic and precision recall curves. The method is tested on the whole human genome. The results for recognizing both translation initiation sites and stop codons indicated improvements in the rates of both false-negative results (FN) and false-positive results (FP). On an average, for translation initiation site recognition, the false-negative ratio was reduced by 30.2% and the FP ratio decreased by 10.9%. For stop codon prediction, FP were reduced by 41.4% and FN by 31.7%.

Availability and implementation: The source code is licensed under the General Public License and is thus freely available. The datasets and source code can be obtained from http://cib.uco.es/site-recognition.

Contact: npedrajas@uco.es

Received on October 10, 2013; revised on March 12, 2014; accepted on May 27, 2014

1 INTRODUCTION

The recognition of translation initiation sites (TISs) and stop codons (Zien *et al.*, 2000) is one of the most critical tasks for gene structure prediction. Most successful current gene recognizers first implement a step of site recognition (Gross *et al.*, 2007), which is followed by a process of combining the sites into meaningful gene structures. This first step is of the utmost importance because the program cannot find genes whose functional sites are not identified. Furthermore, a large number of -FP results might inundate the second step of the programs, making it difficult to predict accurate gene structures.

The best current approaches use powerful classifiers, namely support vector machines (SVMs), and moderately large sequences around the functional site (Baten *et al.*, 2006; Degroeve *et al.*, 2005; Sonnenburg *et al.*, 2007, Zien *et al.*, 2000). In accordance with common practices in machine learning, these methods construct a positive instance set using sequences that contain true TISs or stop codons and a negative instance set; in the negative instance set, the sequences centered around an ATG triplet are not TISs, and the sequences centered around TAG, TAA or TGA triplets are not stop codons. The negative sequences are obtained from all the available information or are randomly selected when sampling is used (García-Pedrajas *et al.*, 2012). Thus, negative sequences can be part of intergenic regions, introns, exons, UTRs, etc.

However, the negative sequences from these different regions have different features. Therefore, the negative class, which the classifier must learn, is highly non-homogeneous. This inhomogeneity is an unnecessary difficulty that the learning algorithm must face and that might damage its performance. In this article, we show how the performance of the classifier can be actually improved if the negative instances are divided into different classes based on their position in the gene; subsequently, different classifiers are learned for each pair of positive and negative instance sets.

Some previous works have also considered the idea of differentiating between functional sites before proceeding to their recognition. TriTISA (Hu *et al.*, 2009) is a method for detecting TISs in microbial genomes that classifies all candidate TISs into three categories based on evolutionary properties, and characterizes them in terms of Markov models. Also, other methods (Burge and Karlin, 1997) have developed different models depending on the structure and composition of the sequences to recognize. However, these approaches are different from ours, as these models are trained and used separately instead of combined as in our proposal.

2 APPROACH

As explained in the previous section, our approach is based on separating the negative sequences based on their position in the gene. The same methodology was used for TIS and stop

*To whom correspondence should be addressed.

codon recognition. Thus, five different sets are created. First, we create a set containing all sequences that contain positive instances. Then, four additional sets are created containing negative sequences; these sets vary based on the position in the gene of those negative sequences. One set was created for each of the three following types of sequences: exons, introns and intergenic regions. A fourth set of negative sequences was created using sequences from non-coding regions, that is, from introns and intergenic regions together. As stated, the aim of this partitioning of the negative set is to obtain more homogeneous negative sets.

In the second step, we must decide how to use these five sets of instances. A straightforward approach would be to use any classifier that can handle more than two classes. However, as mentioned, SVMs are the best performing classifiers for both TIS and stop codon prediction. Although multi-class methods have been developed for SVMs (Hsu and Lin, 2002), two-class approaches usually outperform those methods (Rifkin and Klautau, 2004). Thus, we chose to train four different classifiers, with each classifier trained to differentiate between the positive class and one of the four different negative classes. This approach has the additional advantage that overwhelming evidence in the machine learning literature indicates that a combination of different learners frequently outperforms methods using only one classifier (Rokach, 2009).

Additionally, we use another method in our approach. This last method is the stop codon method. This method is chosen because it only uses positive sequences; thus, it is not affected by the problem of mixing in the different instances of the negative classes. The stop codon method (Saeys *et al.*, 2007) looks at either the stop codon frequencies downstream of the TIS for TIS recognition or the stop codon frequencies upstream of the actual stop codon for stop codon recognition.

After these two steps, we have five trained classifiers that must be combined to obtain a single value that tells us whether a certain sequence is a true site. To classify a new sequence, we obtain the output of each classifier and use those five outputs to predict the class of the sequence. There are many ways of combining the outputs of different classifiers (Kuncheva, 2002), some with high complexity. However, in most cases, simple methods are not beaten by the most complex ones, and these simple methods are faster and less prone to over-fitting. The most common of these simple methods include the sum of the outputs, majority voting and the maximum output.

Given the outputs of the five classifiers, $c_1, \ldots, c_5$, and a threshold for each one of these classifiers, $t_1, \ldots, t_5$, the final answer of the classifier, $C(\mathbf{x})$, for a given sequence x, is defined as follows. For the sum of outputs, the final answer is given by the following equation:

$$C(\mathbf{x}) = \sum_i c_i(\mathbf{x}) - t_i. \tag{1}$$

The threshold term, t_i, corrects for the different ranges of the classifiers. The majority voting approach is given by the following equation:

$$C(\mathbf{x}) = \arg \max_{y \in \{-1,+1\}} \sum_{y : c_i(\mathbf{x}) = y} 1. \tag{2}$$

Finally, the maximum is given by the following equation:

$$C(\mathbf{x}) = c_i(\mathbf{x}) : i = \arg \max_j |c_j(\mathbf{x}) - t_j|. \tag{3}$$

Once $C(\mathbf{x})$ is obtained by any of these methods, a general threshold T should be fixed to decide whether a certain sequence is an actual site. One of the problems we have when choosing the combination method is model selection, as we do not know a priori whether any of the three methods would be consistently better than the other two for all the chromosomes and all the three evaluation measures we used as performance measures. Thus, the best combination was chosen for each case using cross-validation. This cross-validation method is explained in the next section.

As a final remark, we should note that our approach is also general enough to be used with any other classifier. Because this approach is based on modifying the number of classifiers and the training sets, it can be used with any other classification method. Furthermore, this method can also be applied if a classifier uses other types of data besides the raw sequence if the information used by the classifier is extracted using the datasets described above.

3 METHODS AND DATA

To evaluate our approach, we chose five different human chromosomes, namely chromosomes 1, 3, 13, 19 and 21 for testing purposes, and chromosome 16 for model selection. For each chromosome, we trained the classifiers with all the remaining chromosomes except 16, then we chose the best combination method using chromosome 16 and tested the chosen model with all the true TIS or stop codons and the negative samples of the given chromosome. That is, for chromosome 1, we trained the models with chromosomes 2 to 22 and X and Y except 16. Then, we chose the best combination method using chromosome 16 and tested this model using chromosome 1. A summary of these datasets is shown in Table 1. The chromosomes were selected with the aim of choosing chromosomes of different lengths and codification density. Chromosome 16 was chosen as validation set, as it is a chromosome of average length and coding density. For SVMWD, as no model selection is needed, chromosome 16 was added to the training set. We used all TISs and stop codons of the Consensus CDS (CCDS) update released for human of September 7, 2011. This update uses Human National Center for Biotechnology Information (NCBI) build 37.3 and includes 26 473 CCDS IDs that correspond to 18 471 GeneIDs.

One of the key aspects of the evaluation of any new proposal is the set of previous methods used in the comparison. Many different methods have been proposed for recognizing TISs and stop codons (Saeys *et al.*, 2007; Wang *et al.*, 2003; Zeng *et al.*, 2002, Zien *et al.*, 2000). However, these previous works and our own research (García-Pedrajas *et al.*, 2012) have shown that an SVM with a string kernel is the best state-of-the-art method not only for TISs and stop codons but also for splice sites (Sonnenburg *et al.*, 2007). To assure the general advantage of SVMs with string kernels, we performed a preliminary study of the different available methods that included position weight matrices, decision trees, *k*-nearest neighbors, stop codon method (Saeys *et al.*, 2007), Wang *et al.*'s method (Wang *et al.*, 2003), Salzberg's method (Salzberg, 1997) and SVMs with linear and Gaussian kernels and three different string kernels: the locality improved kernel, the weighted degree kernel (WD) and the weighted degree kernel with shifts (Rätsch *et al.*, 2005) (WDS). SVMs with WD kernel obtained consistently the best results and thus was chosen as the method to be compared with our proposal. WDS obtained marginally better results than WD but with a far higher

Table 1. Summary of the training and testing sets

Dataset		Training data Positives/negatives	Testing data Positives	Negatives
Chr. 1	TIS	17 638	2156	8 074 590
	STOP	17 404	2154	23 573 031
Chr. 3	TIS	18 631	1163	7 291 951
	STOP	18 444	1114	21 522 500
Chr. 13	TIS	19 454	340	3 664 164
	STOP	19 225	333	10 878 302
Chr. 19	TIS	18 383	1411	1 698 891
	STOP	18 136	1422	4 665 804
Chr. 21	TIS	19 561	233	1 303 634
	STOP	19 558	237	3 726 959

Random undersampling was used for training; thus, the number of negative instances was equal to the number of positive instances.

Table 2. Summary of the window cross-validation

Data (chr)	Positives versus negatives in					
	All	Exons	Introns	Intergenic regions	Non-coding regions	Stop codon
TIS						
1	[–50,50]	[–50, 50]	[–50, 50]	[–50, 50]	[–50, 50]	[0, 500]
3	[–25,75]	[–50, 50]	[–25, 75]	[–25, 75]	[–25, 75]	[0, 500]
13	[–50,50]	[–50, 50]	[–10, 40]	[–10, 40]	[–10, 40]	[0, 500]
19	[–25,75]	[–50, 50]	[–25, 75]	[–25, 75]	[–25, 75]	[0, 500]
21	[–50,50]	[–50, 50]	[–25, 75]	[–10, 40]	[–25, 75]	[0, 500]
STOP						
All	[–90,10]	[–90, 10]	[–90, 10]	[–90, 10]	[–90, 10]	[–500, 0]

The window obtained around the functional site is shown for each classifier.

computational complexity. We will refer throughout the article to the SVM with WD kernels as SVMWD. The same WD kernel was used for the classifiers in our proposal. However, we must bear in mind that our method, as it works on the design of the datasets, can be used with any other classification method.

Another key parameter of the learning process is the window around the functional site that is used to train the classifiers. A further advantage of our approach is that it allows the use of a suitable window for each type of sequence. The value of the window for each classifier was obtained by cross-validation. We, considering the site as offset 0, and did not count the TIS or the stop codon, and we tested the performance of the following windows: [–100, 0], [–75, 25], [–50, 0], [–50, 50], [–25, 0], [–25, 25], [–25, 75], [–10, 15], [–10, 40], [–10, 90], [0, 25], [0, 50] and [0, 100]. For each trained classifier, the best window was chosen. Table 2 shows the window obtained by cross-validation for all the classifiers. For the stop codon method, we used the additional window values of [0, 200], [0, 300], [0, 400] and [0, 500] for TIS recognition and the window values of [–200, 0], [–300, 0], [–400, 0] and [–500, 0] for stop codon recognition.

Table 2 shows interesting results. First, the window for TIS recognition depended on the classifier and the chromosome. However, the window for stop codon prediction was the same for all cases with only one exception. Second, this table also shows that the different classifiers used for TIS recognition had different values; this finding supports our previous claim that using different classifiers has the advantage of allowing better fine-tuning of the learning parameters.

Furthermore, SVMs are sensitive to the learning parameters; thus, we also performed a cross-validation to obtain their values. The WD kernel has two parameters: the standard C parameter of any SVM and the window width of the string kernel. We tested values of 1, 10, 100 and 1000 for C and 12 and 24 for the window width. All eight combinations were evaluated using 10-fold cross-validation, and the best one was chosen. Although it may be argued that this method might result in suboptimal parameters, this method is a good compromise between the performance of the SVM and the high computational cost of evaluating each set of parameters. This same procedure was used for both SVMWD and our approach.

For training the models, we used random undersampling (Hulse *et al.*, 2010) because previous studies have shown its usefulness for TIS recognition (García-Pedrajas *et al.*, 2012). For random undersampling, we used a ratio of 1, which means that the majority class was randomly undersampled until both classes had the same number of instances.

To evaluate the obtained classifiers, we used the standard measures for imbalanced data. Given the number of true-positive results (TP), false-positive results (FP), true-negative results (TN) and false-negative results (FN), we used the sensitivity, $Sn = \frac{TP}{TP+FN}$, and the specificity, $Sp = \frac{TN}{TN+FP}$. The geometric mean of these two measures, $G-\text{mean} = \sqrt{Sp \cdot Sn}$, will be our first classification metric. As a second measure, we used the area under the receiver operating characteristic curve (auROC). However, auROC is independent of class ratios and it can be less meaningful when we have unbalanced datasets (Sonnenburg *et al.*, 2007). In such cases, area under the precision recall curve (auPRC) can be used. This measure is specially relevant if we are mainly interested in the positive class. However, it can be sensible for subsampling. In our results, we use all the positive and negative instances for each one of the five chromosomes tested, so no subsampling is used. This also yields to small auPRC values.

We use these three metrics because they provide two different views of the performance of the classifiers. The auROC and auPRC values describe the general behavior of the classifier. However, when used in practice, we must establish a threshold for the classification of a query pattern. *G*-mean provides the required snapshot of the performance of the classifier when we set the needed threshold.

4 RESULTS AND DISCUSSION

The first step of our experiments was devoted to studying the usefulness of the five different classifiers that we considered. As stated, we have five classifiers that are trained with the same positive class, and a negative class consisting on negative instances from exons, introns, intergenic regions and non-coding regions. We had a fifth classifier using only positive instances. Thus, we tested the performance, as measured by the auROC, of the combined approach with all five classifiers and then removed one classifier at a time. A negative value means that the classifier had a negative effect on the performance of the model and thus should not be used.

The results showed that the worst performing classifier was the one trained using negative instances extracted from exons. For this classifier, the positive and negative instances were the most similar; thus, the training algorithm had more difficulties in differentiating between the positive and negative instances. In fact, the overall effect of this classifier was harmful to the performance of the method. Thus, this classifier was removed and was not considered in the subsequent experiments.

Table 3. Summary of the results for TIS recognition

Dataset	SVMWD							Proposed approach						
	Sp	Sn	TP	FN	TN	FP	auROC/PRC	Sp	Sn	TP	FN	TN	FP	auROC/PRC
Chr. 1	.9155	.8326	1795	361	7392644	681946	.9481/.1001	.9209	.9003	1941	215	7435495	639095	.9693/.1351
Chr. 3	.9024	.8203	954	209	6580426	711525	.9357/.0891	.9066	.9003	1047	116	6611176	680775	.9628/.1229
Chr. 13	.9398	.8294	282	58	3443428	220736	.9522/.0818	.9457	.8824	300	40	3465327	198837	.9695/.1207
Chr. 19	.8961	.8703	1228	183	1522340	176551	.9522/.1358	.9077	.8824	1245	166	1542012	156879	.9551/.1321
Chr. 21	.8965	.8326	194	39	1168732	134902	.9387/.0689	.9200	.8755	204	29	1199340	104294	.9691/.1203

The table shows the specificity (Sp), sensitivity (Sn),TP, TN, FN, FP and area under the ROC and PRC curves (auROC/PRC) for both methods and the five studied chromosomes.

The results also showed that the stop codon method classifier had the most important contribution. This finding is interesting because this classifier was the worst when considered alone. The explanation for this difference may be found in the behavior of the ensembles of classifiers. It is well known (Kuncheva and Whitaker, 2003) that a diverse ensemble of classifiers improves the performance of the set of classifiers. The stop codon method differs from the other four classifiers, which are all based on SVMs; thus, although its performance is worse than the performance of those four classifiers individually, the diversity it introduces improves the performance of the set of classifiers.

The next step was the comparison of the performances of our approach and SVMWD. A summary of the results for TIS recognition of the five studied chromosomes is shown in Table 3. The first interesting result is that the proposed approach beat SVMWD for all measures and all chromosomes with only one exception. The improvements in specificity, sensitivity, geometric mean, auROC and auPRC are shown in Figure 1.

The second remarkable result shown in Table 3 is the significant reduction in the FN rate. The reduction in the number of FN was 9.3% in the worst case and 44.5% in the best case. This reduction means that 284 TISs that were inaccurately classified as negatives by SVMWD were correctly identified by our method. Most current gene recognizers rely heavily on the classification of TISs; therefore, it is likely that those genes would be completely missed by any gene recognizer. Thus, our approach has the potential to improve the accuracy of any annotation system by 6.4%.

Furthermore, our method was also able to improve the true-negative rate. In total, 145 780 FP from SVMWD were correctly classified as negatives using our approach. Therefore, any annotation system that uses our metric would have a significantly reduced set of putative TISs and better expected performance.

The improvement for auROC and auPRC values are also shown in Figure 1 (We always performed the testing of all the methods with all the negative samples. That means that the ratio minority/majority class is almost 1:11 000 for the worst case yielding to low auPRC values. We must take into account that with only a few thousand FP among several millions of TN we would obtain a low precision value. The situation for stop codon recognition is even worse as the number of TN is multiplied by three). The actual ROC and PRC curves are shown in Figures 2–6. These figures show that our approach improved the auROC

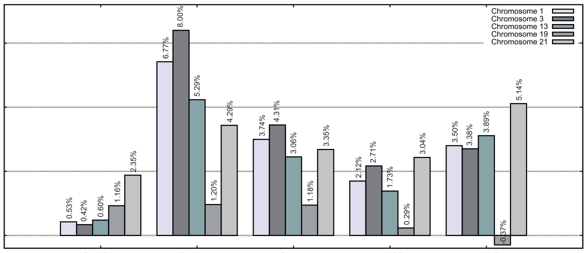

Fig. 1. Absolute improvement for TIS recognition in specificity, sensitivity, *G*-mean, auROC and auPRC of our approach compared with SVMWD

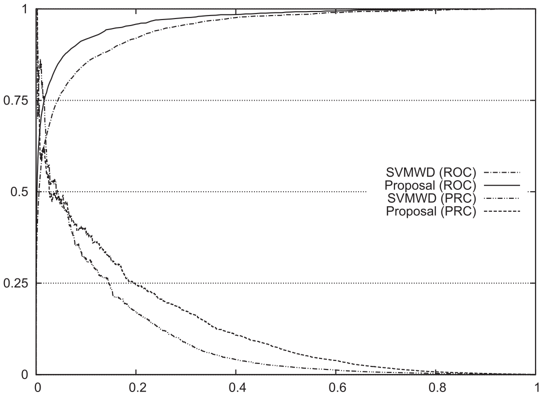

Fig. 2. ROC/PRC curves for TIS prediction for chromosome 1

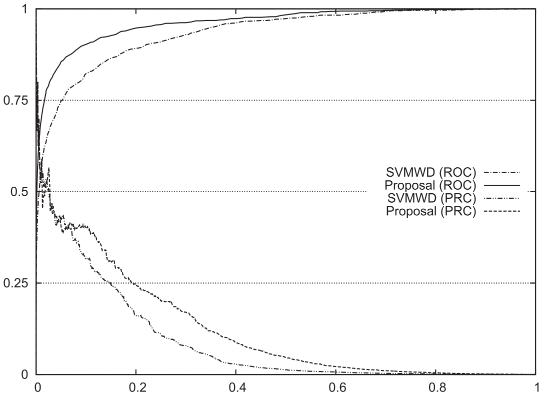

Fig. 3. ROC/PRC curves for TIS prediction for chromosome 3

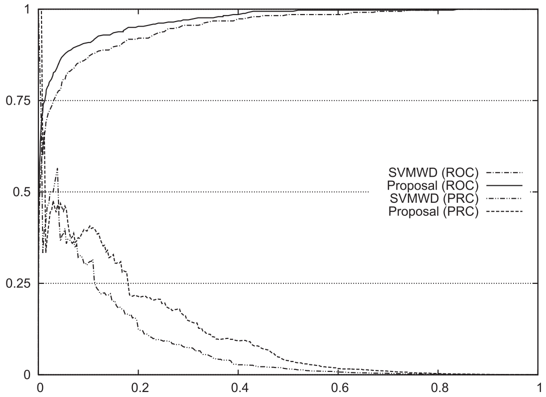

Fig. 4. ROC/PRC curves for TIS prediction for chromosome 13

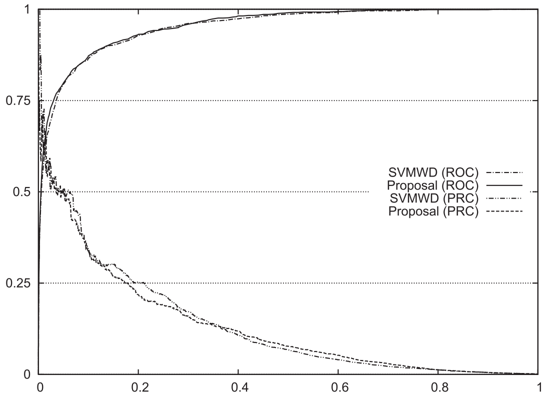

Fig. 5. ROC/PRC curves for TIS prediction for chromosome 19

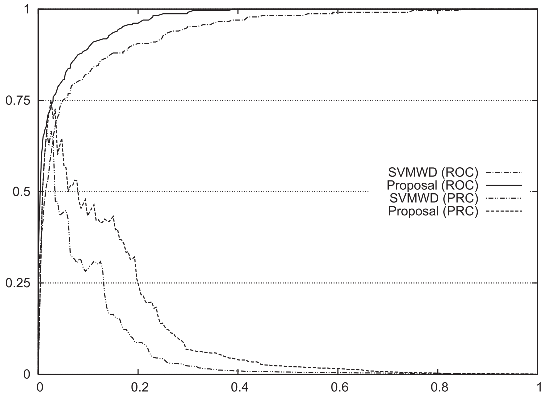

Fig. 6. ROC/PRC curves for TIS prediction for chromosome 21

and auPRC for all five studied chromosomes. These results demonstrate that the overall performance of the proposed method was better than the performance of SVMWD. The actual ROC and PRC curves shown in Figures 2–6 show that the curves corresponding to our proposal are always above than the curves of SVMWD. This result indicates the better performance for all the possible thresholds of classification.

It is interesting to study how the proposed method achieved its good performance. For both methods, Table 4 shows the distribution of the FP according to the part of the gene to which the TIS sequences belong. The behavior is clear.

Table 4. Distribution of FP for both methods

Dataset	SVMWD			Proposed approach		
	Exon	Intron	Intergenic regions	Exon	Intron	Intergenic regions
Chromosome 1	0.15%	1.39%	6.91%	0.21%	0.71%	3.72%
Chromosome 3	0.12%	1.58%	8.06%	0.13%	0.56%	4.24%
Chromosome 13	0.06%	0.79%	5.17%	0.07%	0.37%	2.79%
Chromosome 19	0.58%	1.55%	8.26%	0.71%	1.23%	6.54%
Chromosome 21	0.10%	1.30%	8.95%	0.13%	0.63%	4.05%
Average	0.20%	1.32%	7.47%	0.25%	0.70%	4.27%

The table shows the type of genome region for each of the FP

Separating the negatives samples into four classes improves the discrimination between positive instances and negative instances from introns and intergenic regions. However, the number of FP for instances from exons increases but to a lesser extent than the decrease in the number of FP from introns and intergenic regions. Furthermore, the FP from exons may be reduced using other sources of information, such as content measures.

The second part of our experiments was devoted to stop codon recognition. Stop codon recognition is a more difficult task because the achieved accuracy is less than that for TIS recognition. One of the major sources of this increased complexity is the number of negative instances. There are three different stop codons rather than just one as it is the case for TIS recognition; therefore, the number of negative instances is three times the number of negative instances for TIS prediction. For instance, using the same five chromosomes from the previous experiments, the best current method found >11.5 million false-positive stop codons. This amount of incorrectly predicted stop codons might be able to mar any annotation system, indicating that there is ample room for improvement.

Our approach for stop codon prediction used the same classifiers as for TIS recognition. Table 5 shows a summary of results for stop codon recognition. The first remarkable result is the large improvement in the number of both FN and FP. The FN were reduced by 26.8% in the worst case and by 57.9% in the best case. This result indicates that the number of total FN was reduced from 823 with SVMWD to 443 with our method. As for TIS recognition, an annotation program may not be able to recognize a gene when the stop codon is missing. Furthermore, this improvement was achieved along with a significant improvement in the FP. The FP improvement was also large with a best result of 46.9%. As a whole, 2.7 million FP from SVMWD were accurately classified as negatives by our method. This quantity of FP may overwhelm any annotation system; thus, the improvement should have a significant impact on automatic annotation.

Figure 7 shows the absolute improvement of our method in the specificity, sensitivity, *G*-mean, auROC and auPRC. The improvement for auROC is particularly relevant. The proposed approach improved the auROC by 3.7% in the worst case and 7.2% in the best case. Figures 8–12 show the ROC/PRC curves for the five chromosomes. As for TIS recognition, the ROC/PRC curves of our approach not only achieved a better auROC and auPRC but were also always above the curves of SVMWD.

Table 5. Summary of the results for STOP codon recognition

Dataset	SVMWD							Proposed approach						
	Sp	Sn	TP	FN	TN	FP	auROC/PRC	Sp	Sn	TP	FN	TN	FP	auROC/PRC
Chr. 1	.8361	.8347	1798	356	19710568	3862463	.9182/.0127	.8393	.9304	2004	150	19784595	3788436	.9583/.0209
Chr. 3	.8109	.8402	936	178	17452195	4070305	.9115/.0049	.8630	.9174	1022	92	18573710	2948790	.9584/.0133
Chr. 13	.8183	.8318	277	56	8901404	1976898	.9073/.0067	.8900	.8769	292	41	9681647	1196655	.9494/.0081
Chr. 19	.8117	.8706	1238	184	3787466	878338	.9258/.0357	.9000	.9058	1288	134	4199221	466583	.9630/.0553
Chr. 21	.8089	.7932	188	49	3014762	712197	.8811/.0058	.8900	.8903	211	26	3316990	409969	.9533/.0132

The table shows the specificity (Sp), sensitivity (Sn), TP, TN, FN, FP and area under the ROC and PRC curves (auROC/PRC) for both methods and the five studied chromosomes.

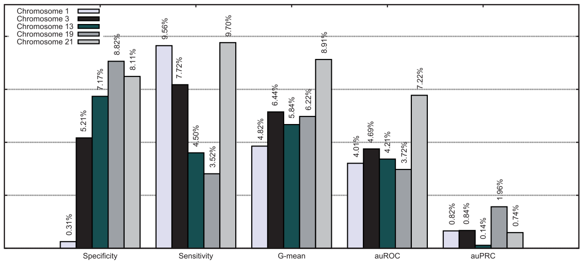

Fig. 7. Absolute improvement for stop codon recognition in the specificity, sensitivity and *G*-mean of our approach compared with SVMWD

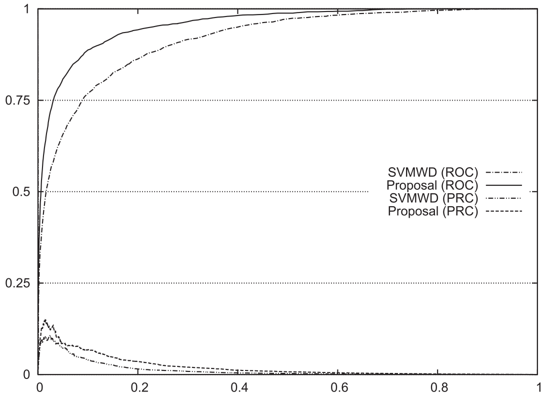

Fig. 8. ROC/PRC curves for stop codon prediction for chromosome 1

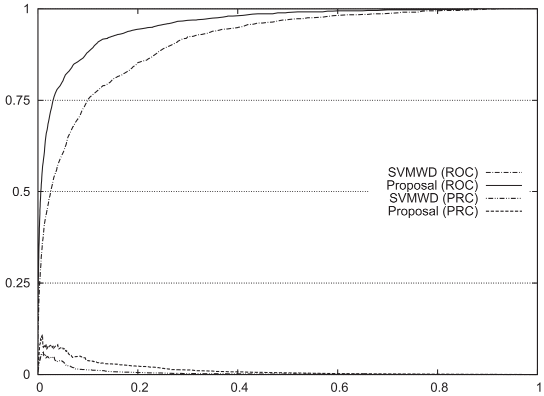

Fig. 9. ROC/PRC curves for stop codon prediction for chromosome 3

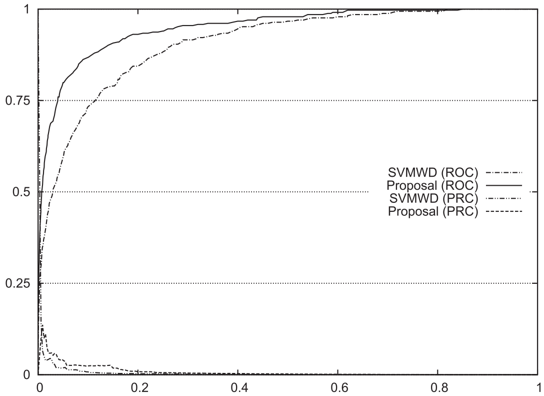

Fig. 10. ROC/PRC curves for stop codon prediction for chromosome 13

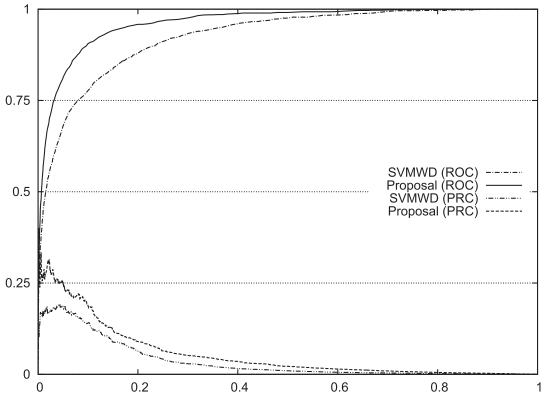

Fig. 11. ROC/PRC curves for stop codon prediction for chromosome 19

In Section 2, we stated that our approach could be applied to any type of classifier. In the previous experiments, we used SVMs because they achieved the best performance in the literature. Now, we present the results of another experiment that was conducted to demonstrate the applicability of our method to other classifiers. We used a decision tree using the C4.5 learning algorithm (Quinlan, 1996) instead of SVMs. We tested a decision tree using the standard approach of only one training set and one classifier and using our method with the four classifiers that were used for SVMs. To avoid repeating all the experiments, we only performed experiments for chromosome 13. The results for both TIS and stop codon recognition are shown in Table 6. For TIS

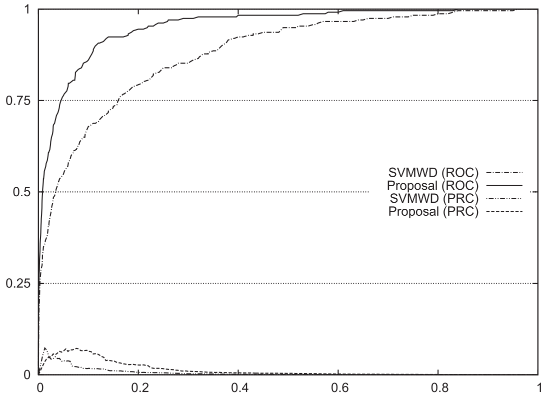

Fig. 12. ROC/PRC curves for stop codon prediction for chromosome 21

Table 6. Results for TIS and stop codon prediction for chromosome 13 using a decision tree as the classifier

Dataset	Method	auROC/auPRC	Sp	Sn	*G*-mean
TIS	C4.5	0.8183/0.0005	0.7756	0.7706	0.7731
	Proposed	0.9372/0.0154	0.9053	0.8000	0.8510
Stop codon	C4.5	0.6853/0.0001	0.6489	0.6426	0.6458
	Proposed	0.9260/0.0097	0.8468	0.8709	0.8587

The table shows the values of the specificity (Sp), sensitivity (Sn), geometric mean of the specificity and sensitivity and the area under the ROC/PRC curves (auROC/auPRC).

recognition, the improvement was remarkable; the *G*-mean improved by 8%, and the auROC increased from 0.8189 to 0.9372. For stop codon classification, the improvement was even better. The standard approach had an auROC of 0.6853, whereas our approach achieved an auROC of 0.9260. As it was the case for the previous experiments, auPRC was low for all experiments because of the huge number of negative instances.

5 CONCLUSION

In this article, we presented a new approach for TIS and stop codon recognition. This approach uses more than one classifier, divides the negative class into four different groups and trains one classifier for each type of negative class. This approach was applied to the recognition of TIS and stop codons in five human chromosomes. The approach was compared with the best current method for TIS and stop codon prediction. The proposed approach also has the advantage of its simplicity, which makes it easily applicable to any program for TIS or stop codon recognition.

The reported results show that the proposed method shows improved sensitivity, specificity, auROC and auPRC compared with SVMWD. The results show a remarkable improvement in the ratio of FN and FP achieved over those of SVMWD. Because state-of-the-art annotation systems rely heavily on the accurate prediction of the functional sites of the gene, the proposed method is an effective way of improving current gene recognizers.

Funding: This work has been financed in part by Project TIN-2011-22967 of the Spanish Ministry of Science and Innovation and Excellence in Research Projects P09-TIC-4623 and P07-TIC-2682 of the Junta de Andalucía.

Conflict of Interest: none declared.

REFERENCES

Baten,A. *et al.* (2006) Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics*, **7** (**Suppl. 5**), S15.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Degroeve,S. *et al.* (2005) SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics*, **21**, 1332–1338.

García-Pedrajas,N. *et al.* (2012) Class imbalance methods for translation initiation site recognition in DNA sequences. *Knowl. Based Syst.*, **25**, 22–34.

Gross,S.S. *et al.* (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.*, **8**, R269.

Hsu,C.W. and Lin,C.J. (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.*, **13**, 415–425.

Hu,G.Q. *et al.* (2009) Prediction of translation initiation site for microbial genomes with tritisa. *Bioinformatics*, **25**, 123–125.

Hulse,J.V. *et al.* (2010) An empirical evaluation of repetitive undersampling techniques. *Int. J. Softw. Eng. Know. Eng.*, **20**, 173–195.

Kuncheva,L. (2002) A theoretical study of six classifier fusion strategies. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 281–286.

Kuncheva,L. and Whitaker,C.J. (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, **51**, 181–207.

Quinlan,J.R. (1996) Bagging, boosting, and c4.5. In: *Proceedings if the Thirteenth National Conference on Artificial Intelligence*. AAAI Press and the MIT Press, pp. 725–730.

Rätsch,G. *et al.* (2005) RASE: recognition of alternative spliced exons in *c. elegans*. *Bioinformatics*, **21** (**Suppl. 1**), i369–i377.

Rifkin,R. and Klautau,A. (2004) In defense of one-vs-all classification. *J. Mach. Learn. Res.*, **5**, 101–141.

Rokach,L. (2009) Taxonomy for characterizing ensemble methods in classification tasks: a review and annotated bibliography. *Comput. Stat. Data Anal.*, **53**, 4046–4072.

Saeys,Y. *et al.* (2007) Translation initiation site prediction on a genomic scale: beauty in simplicity. *Bioinformatics*, **23**, 418–423.

Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci*, **13**, 365–376.

Sonnenburg,S. *et al.* (2007) Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, **8** (**Suppl. 10**), S7.

Wang,Y. *et al.* (2003) Recognizing translation initiation sites of eukaryotic genes based on the cooperatively scanning model. *Bioinformatics*, **19**, 1972–1977.

Zeng,F. *et al.* (2002) Using feature generation and feature selection for accurate prediction of translation initiation sites. *Genome Inform.*, **13**, 192–200.

Zien,A. *et al.* (2000) Engineering support vector machines kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.